Случайные процессы ПМИ

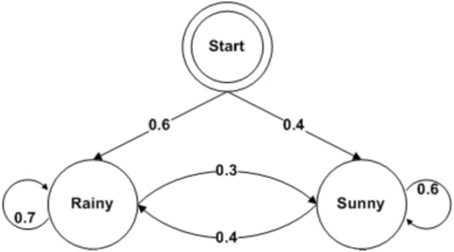
Прикладной поток

Семинар 4

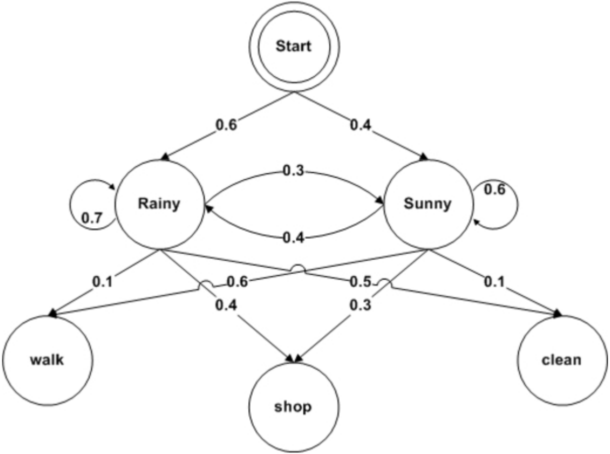
ФИВТ МФТИ

1. Скрытые марковские модели

Марковская цепь



Скрытая марковская модель



Определение

Пусть $X = (X_n, n \in \mathbb{Z}_+)$ — однородная марковская цепь со значениями в $\mathscr{X} = \{1, ..., r\}$ с матрицей переходных вероятностей $P = (p_{ij})_{i,j=1}^r$ и стационарным распределением $\Pi = (\pi_j)_{j=1}^r$, причем X_n ненаблюдаемы (называются скрытыми).

Пусть $Y = (Y_n, n \in \mathbb{Z}_+)$ — наблюдаемый случайный процесс, такой что условная плотность вектора $Y_1, ..., Y_T$ равна

$$f_{Y_1,...,Y_T|X_1,...,X_T}(y_1, ..., y_T|x_1, ..., x_T) = \prod_{j=1}^{T} f_{x_j}(y_j),$$

где $f_1, ..., f_r$ — либо настоящие плотности, либо дискретные.

Пара процессов (X, Y) называется **скрытой марковской моделью**.

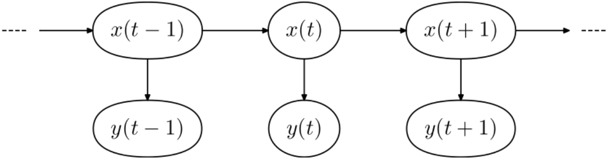
Следствие

Тем самым,

1. $Y_1, \ldots, Y_T$ — условно независимы, то есть Y_j непосредственно зависит только от X_j.
2. Плотность каждого Y_i равна
$$f(y) = \sum_{j=1}^{r} \mathrm{P}(X_i = j) f_j(y) = \sum_{j=1}^{r} \pi_j f_j(y).$$

Частный случай — величины X_n независимы. Это модель смеси распределений, см. задачу про Ирисы Фишера из статистики.

Скрытая марковская модель



Предположения модели

1. Все $f_1, ..., f_r$ лежат в некотором параметрическом семействе $\{f(\theta), \theta \in \Theta\}, \Theta \subset \mathbb{R}, f(y) = \sum_{j=1}^{n} \pi_j f(y, \theta_j)$.
2. Неизвестными параметрами являются $\psi = (\theta_i, \pi_i, p_{ij}, i, j = 1, ..., r)$, $\theta_i = \theta_i(\psi), \pi_i = \pi_i(\psi), p_{ij} = p_{ij}(\psi)$. И пусть ψ_0 — истинное значение.

Задачи:

1. оценить параметры с помощью ОМП.
2. оценить значения скрытых состояний X_n.

2. Приложения

Детектор лжи

По вибрации рук человека, определяет, говорит он правду или нет.

Когда человек лжет, руки трясутся чуть больше.

Детектор лжи

$\mathscr{X} = \{True, False\}$ — множество состояний

$X = (X_1, ..., X_T)$ — последовательность скрытых состояний, где X_n — правда или ложь

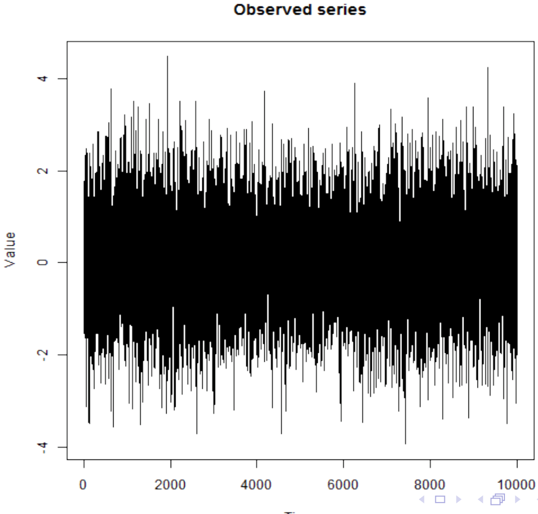
$Y = (Y_1, ..., Y_T)$ — сигнал с устройства

f_{true} — плотность распределения $\mathcal{N}(0, \sigma_1^2)$, а f_{false} — плотность распределения $\mathcal{N}(0, \sigma_2^2)$, причем $\sigma_1^2 < \sigma_2^2$.

В этом случае матрица P имеет большие значения на диагонали.

Детектор лжи

Сигнал с устройства:

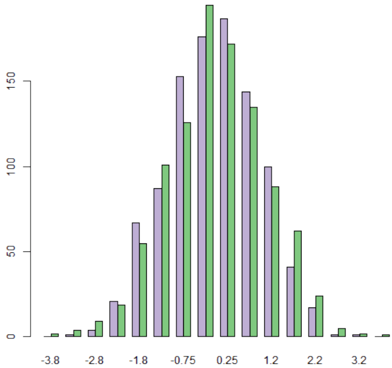


Детектор лжи

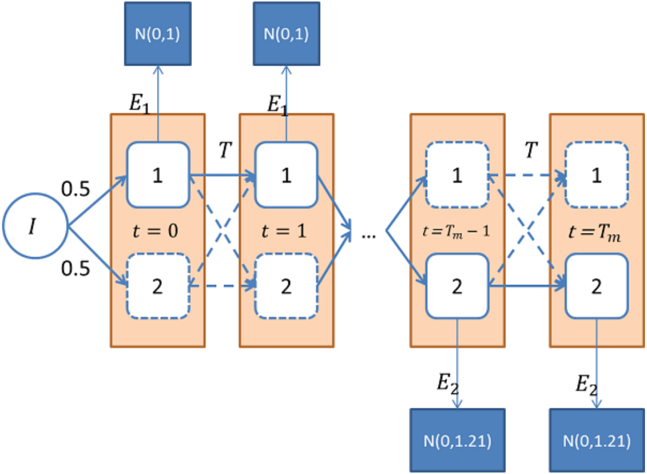
Гистограмма состояний:

правда — $\mathcal{N}(0, 1)$

ложь — $\mathcal{N}(0, 1.21)$

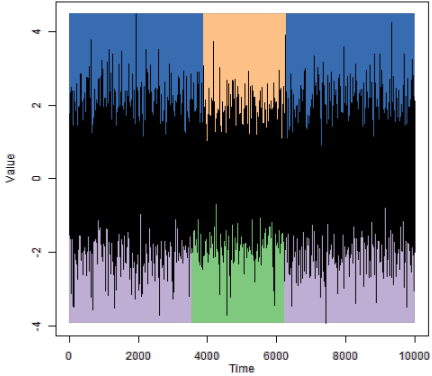


Детектор лжи



Детектор лжи

Результат применения HMM:



Оценка: правда, ложь; Истинные значения: **правда**, **ложь**

Подробнее: https://habrahabr.ru/post/180109/

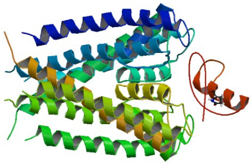
Биоинформатика

Y — аминокислотная последовательность трансмембранного белка.

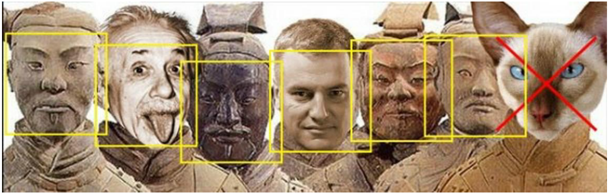
Белок состоит из частей двух типов $\mathscr{X} = \{\text{трансмембранная}, \text{растворимая}\}$.

Известно, что частоты встречаемости аминокислот в трансмембранных и в растворимых частях белка различаются.

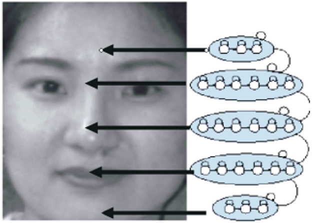
Определить по последовательности где находятся трансмембранные участки.



Поиск лиц



HMM для поиска лиц состоит из 5 супер-состояний, соответствующих областям лица (лобовая часть, глаза, нос, рот, подбородок), каждое из которых делится на отдельные состояния.



Подробнее: https://habrahabr.ru/post/109956/

Part-of-speech tagging

Задача: выбрать правильный морфологический разбор.

Хранение денег в **банке**.

Что делают **белки** в клетке?

Фотографии **Львов**.

Капля **стекла** со **стекла**.

Косил **косой косой косой**.

Полосы **стали** красными...

Полосы **стали** красными реками текли по конвейеру трубопрокатного завода.

Penn Treebank

CC	Coordinating conjunction	NNS	Noun, plural	TO	to
CD	Cardinal number	NNP	Proper noun, singular	UH	Interjection
DT	Determiner	NNPS	Proper noun, plural	VB	Verb, base form
EX	Existentialthere	PDT	Predeterminer	VBD	Verb, past tense
FW	Foreign word	POS	Possessive ending	VBG	Verb, gerund or present participle
IN	Preposition or subordinating conjunction	PRP	Personal pronoun	VBN	Verb, past participle
JJ	Adjective	PRP$	Possessive pronoun	VBP	Verb, non-3rd person singular present
JJR	Adjective, comparative	RB	Adverb	VBZ	Verb, 3rd person singular present
JJS	Adjective, superlative	RBR	Adverb, comparative	WDT	Wh-determiner
LS	List item marker	RBS	Adverb, superlative	WP	Wh-pronoun
MD	Modal	RP	Particle	WP$	Possessive wh-pronoun
NN	Noun, singular or mass	SYM	Symbol	WRB	Wh-adverb

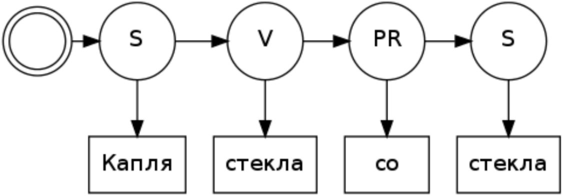
Part-of-speech tagging

$\mathscr{X}$ — множество тегов

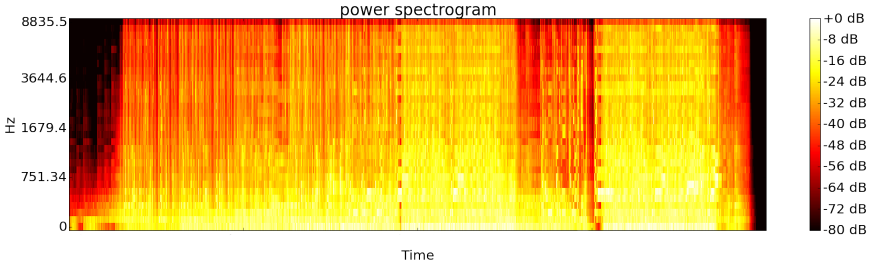
$X = (X_1, ..., X_T)$ — последовательность тегов предложения

$Y = (Y_1, ..., Y_T)$ — слова предложения

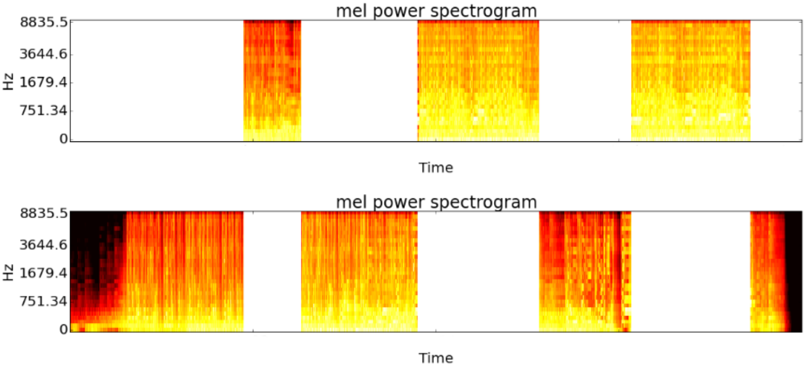
Появление очередного слова текста будет зависеть от текущего морфологического тега, а появление очередного тега — от предыдущих тегов



Музыкальный трек



Сегментация музыкального трека



3. Оценка параметров

Свойства ОМП

Условия:

- $P(\psi_0)$ эргодична (для нее выполнены условия эргодической теоремы) (следовательно, π определяется по P).
- Семейство $\{f(y,\theta), \theta \in \Theta\}$ идентифицируемо, то есть сумма

  $$f(y) = \sum_{j=1}^{r} \pi_j f(y, \theta_j)$$

  представима однозначно, по ней наборы $(\pi_1, ..., \pi_r)$ и $(\theta_1, ..., \theta_r)$ определяются с точностью до перестановки при различных $\theta_1, ..., \theta_r$.
- $f(y,\theta)$ непрерывна по θ и $\Theta \subset \mathbb{R}^k$ — открытое множество.
- Функции $p_{ij}(\psi)$ и $\theta_j(\psi)$ непрерывны по ψ.

Свойства ОМП

Условия:

- $\mathsf{E}_{\psi_0} |\ln f(Y_1, \theta_j(\psi))| < +\infty \quad \forall j = 1, \dots, r$
- $\forall \theta \in \Theta : \mathsf{E}_{\psi_0} \left(\sup\limits_{\theta' : ||\theta' - \theta|| < \delta} (\ln f(Y_1, \theta))^+ \right) < \infty$ для некоторого $\delta > 0$.
- $\theta_j(\psi_0) \neq \theta_i(\psi_0)$ при $i \neq j$.

Теорема

В условиях 1-7 ОМП сходится по вероятности к некоторому $\widetilde{\psi}$, который совпадает с ψ_0 при перестановке индексов.

Brian G. Leroux "Maximum-likelihood estimation for hidden Markov models"

Поиск ОМП

Введем некоторые обозначения

1. $X = (X_1, ..., X_T), Y = (Y_1, ..., Y_T), x = (x_1, ..., x_T), y = (y_1, ..., y_T)$
2. Вероятность быть в состоянии k в момент времени t при условии наблюдаемой последовательности $Y = y$
$$L_k(t) = \mathsf{P}(X_t = k | Y = y)$$
3. Вероятность быть в состоянии k в момент времени t и перейти в состоянии l в следующий момент времени при условии наблюдаемой последовательности $Y = y$
$$H_{kl}(t) = \mathsf{P}(X_t = k, X_{t+1} = l | Y = y)$$

Заметим, что $L_k(t) = \sum\limits_{l=1}^{r} H_{kl}(t), \sum\limits_{k=1}^{r} L_k(t) = 1.$

EM-алгоритм

Максимизация функции правдоподобия происходит с помощью итерационного EM-алгоритма, который заключается в чередовании E-шага и M-шага.

E-шаг: для текущего приближения параметров вычислить значения вероятностей $L_k(t)$ и $H_{kl}(t)$.

M-шаг:

$$p_{kl} = \sum_{t=1}^{T-1} H_{kl}(t) \Bigg/ \sum_{t=1}^{T-1} L_k(t).$$

$$\pi_k = \sum_{t=1}^{T} L_k(t) \Bigg/ \sum_{k=1}^{r} \sum_{t=1}^{T-1} L_k(t).$$

+ обновить параметры распределений f_k.

EM-алгоритм

Обновление параметров распределений f_k:

1. Если f_k — плотность распределения $\mathcal{N}(a_k, \Sigma_k)$, то

$$a_k = \sum_{t=1}^{T} L_k(t) Y_t \Big/ \sum_{t=1}^{T} L_k(t),$$

$$\Sigma_k = \sum_{t=1}^{T} L_k(t)(Y_t - a_k)(Y_t - a_k)^T \Big/ \sum_{t=1}^{T} L_k(t).$$

2. Если $f_k(j) = q_{kj}$ — дискретная плотность, то

$$q_{kj} = \sum_{t=1}^{T} L_k(t) I\{Y_t = j\} \Big/ \sum_{t=1}^{T} L_k(t).$$

EM-алгоритм

- Чередование шагов начинается с некоторой начальной инициализацией параметров и до сходимости величины

$$F(\psi) = \sum_{k=1}^{r} L_k(1) \log \pi_k + \sum_{t=2}^{T} \sum_{k=1}^{r} \sum_{l=1}^{r} H_{kl}(t) \log p_{kl}$$
$$+ \sum_{t=1}^{T} \sum_{k=1}^{r} L_k(t) \log f_k(y_t, \theta)$$

- Может сходиться к локальному максимуму, поэтому имеет смысл запускать несколько раз из разных начальных приближений.
- Связь со смесью распределений: величины $L_k(t)$ играют ту же роль, что и апостериорная вероятность компоненты при условии наблюдения.

Метод Forward-Backward

Используется для эффективного вычисления $L_k(t)$ и $H_{kl}(t)$.

Теорема

Вместо прямого вычисления $L_k(t)$ и $H_{kl}(t)$, которое требует экспоненциальное количество времени (по r), можно использовать эффективный метод со временем $O(r^2 T)$, а используемая память $O(r^2 T)$.

Метод Forward-Backward

Доказательство теоремы:

Определим прямые вероятности

$\alpha_k(t) = \mathrm{P}(Y_1 = y_1, ..., Y_t = y_t, X_t = k).$

Рекурсивные формулы

$$\alpha_k(1) = \pi_k f_k(y_1),$$

$$\alpha_k(t) = f_k(y_t) \sum_{l=1}^{r} \alpha_l(t-1) p_{lk}.$$

Метод Forward-Backward

$$\alpha_k(t) = \mathrm{P}(Y_1 = y_1, \ldots, Y_t = y_t, X_t = k) =$$

$$= \sum_{l=1}^{r} \mathrm{P}(Y_1 = y_1, \ldots, Y_t = y_t, X_t = k, X_{t-1} = l) =$$

$$= \sum_{l=1}^{r} \mathrm{P}(Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}, X_{t-1} = l) \times$$

$$\times \mathrm{P}(Y_t = y_t, X_t = k | X_{t-1} = l, Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}) =$$

$$= \sum_{l=1}^{r} \alpha_l(t-1)\, \mathrm{P}(Y_t = y_t, X_t = k | X_{t-1} = l) =$$

$$= \sum_{l=1}^{r} \alpha_l(t-1)\, \mathrm{P}(Y_t = y_t | X_t = k, X_{t-1} = l)\, \mathrm{P}(X_t = k | X_{t-1} = l) =$$

$$= \sum_{l=1}^{r} \alpha_l(t-1)\, \mathrm{P}(Y_t = y_t | X_t = k)\, \mathrm{P}(X_t = k | X_{t-1} = l) = \sum_{l=1}^{r} \alpha_l(t-1) f_k(y_t) p_{lk}$$

Метод Forward-Backward

Определим обратные вероятности

$\beta_k(t) = \mathrm{P}(Y_{t+1} = y_{t+1}, \dots, Y_T = y_T | X_t = k)$ и $\beta_k(T) = 1$.

Рекурсивные формулы

$$\beta_k(T) = 1,$$

$$\beta_k(t) = \sum_{l=1}^{r} p_{kl} f_l(y_{t+1}) \beta_l(t+1).$$

Докажете в ДЗ :)

Метод Forward-Backward

Вероятности $L_k(t)$ и $H_{kl}(t)$ можно вычислить по формулам

$$L_k(t) = \frac{\alpha_k(t)\beta_k(t)}{\mathrm{P}(Y = y)}$$

$$H_{kl}(t) = \frac{p_{kl} f_l(y_{t+1})\alpha_k(t)\beta_l(t+1)}{\mathrm{P}(Y = y)}$$

Докажете в ДЗ :)

Метод Forward-Backward

Кроме того, для любого t справедлива формула

$$\mathrm{P}(Y = y) = \sum_{k=1}^{r} \alpha_k(t)\beta_k(t),$$

$$[\text{ при } t = T\ ] \qquad \mathrm{P}(Y = y) = \sum_{k=1}^{r} \alpha_k(T)\beta_k(T) = \sum_{k=1}^{r} \alpha_k(T).$$

Доказательство

$$\mathrm{P}(Y = y) = \mathrm{P}(Y_1 = y_1, \dots, Y_T = y_T) = \sum_{k=1}^{r} \mathrm{P}(Y_1 = y_1, \dots, Y_T = y_T, X_t = k) =$$

$$= \sum_{k=1}^{r} \mathrm{P}(Y_1 = y_1, \dots, Y_t = y_t, X_t = k) \times$$

$$\times \mathrm{P}(Y_{t+1} = y_{t+1}, \dots, Y_T = y_T | X_t = k, Y_1 = y_1, \dots, Y_t = y_t) =$$

$$= \sum_{k=1}^{r} \alpha_k(t)\,\mathrm{P}(Y_{t+1} = y_{t+1}, \dots, Y_T = y_T | X_t = k) = \sum_{k=1}^{r} \alpha_k(t)\beta_k(t)$$

Метод Forward-Backward

Итог:

$$\alpha_k(t) = f_k(y_t) \sum_{l=1}^{r} \alpha_l(t-1) p_{lk}, \quad \alpha_k(1) = \pi_k f_k(y_1)$$

$$\beta_k(t) = \sum_{l=1}^{r} p_{kl} f_l(y_{t+1}) \beta_l(t+1), \quad \beta_k(T) = 1$$

$$L_k(t) = \frac{\alpha_k(t)\beta_k(t)}{\mathrm{P}(Y=y)}$$

$$H_{kl}(t) = \frac{p_{kl} f_l(y_{t+1}) \alpha_k(t) \beta_l(t+1)}{\mathrm{P}(Y=y)}$$

$$\mathrm{P}(Y=y) = \sum_{k=1}^{r} \alpha_k(T).$$

Время $O(r^2 T)$, память $O(r^2 T)$.

4. Оценка траектории для X_n

Траектория forward-backward:

последовательность $x = (x_1, ..., x_T)$, где

$x_t = \arg\max_k L_k(t) = \arg\max_k \mathsf{P}(X_t = k | Y = y)$.

Траектория Витерби:

$$x^* = \arg\max_x \mathsf{P}(X = x | Y = y).$$

Траектория Витерби

$$x^* = \arg\max_x \mathsf{P}(X = x | Y = y) = \arg\max_x \mathsf{P}(X = x, Y = y) =$$
$$= \arg\max_x \log \mathsf{P}(X = x, Y = y).$$

Теорема

Данную задачу дискретной оптимизации можно решить за время $O(r^2 T)$, используя $O(rT)$ памяти с помощью метода динамического программирования, в то время как полный перебор происходит за время $O(r^T)$.

Метод доказательства теоремы был предложен Эндрю Витерби в 1967 году как метод декодирования свёрточного кода, передаваемого по сетям с наличием шума.

Траектория Витерби

Распишем оптимизируемый функционал

$$G(x) = \log \mathrm{P}(X = x, Y = y) = \log(\pi_{x_1} f_{x_1}(y_1) p_{x_1,x_2} f_{x_2}(y_2) \dots p_{x_{T-1},x_T} f_{x_T}(y_T)) =$$

$$= [\log \pi_{x_1} + \log f_{x_1}(y_1)] + [\log p_{x_1,x_2} + \log f_{x_2}(y_2)] +$$

$$+ \dots + [\log p_{x_{T-1},x_T} + \log f_{x_T}(y_T)].$$

Определим

$$g_1(x_1) = \log \pi_{x_1} + \log f_{x_1}(y_1),$$

$$g_t(x_t, x_{t-1}) = \log p_{x_t,x_{t-1}} + \log f_{x_t}(y_t).$$

Тогда

$$G(x) = g_1(x_1) + \sum_{t=2}^{T} g_t(x_t, x_{t-1}).$$

Траектория Витерби

Определим для данной скрытой марковской модели *граф развертки* $G = (V, E)$:

- $V = V_1 \sqcup \ldots \sqcup V_T$, где V_t — копии множества скрытых состояний
.
- $E = \{(x_{t-1}, x_t) | x_{t-1} \in V_{t-1}, x_t \in V_t\}$

Зададим веса:

- ребру (x_{t-1}, x_t), где $x_{t-1} \in V_{t-1}$, $x_t \in V_t$ соответствует вес $g_t(x_t, x_{t-1})$;
- стартовой вершине $x_1 \in V_1$ соответствует вес $g_1(x_1)$. Можно так же предполагать наличие общей стартовой вершины x_0, а $g_1(x_1)$ — вес ребра (x_0, x_1).

Траектория Витерби

Алгоритм:

1. $t = 1$: для $k = 1, \ldots, r$ записываем $G^*_{1,k} = g_1(k)$
2. $t \geqslant 2$: для $k = 1, \ldots, r$ вычисляем, какое состояние будет оптимальным на предыдущем шаге
$$G^*_{t,k} = \max_{l=1\ldots r} \left(G^*_{t-1,l} + g_t(k,l)\right) = G^*_{t-1,l^*} + g_t(k,l^*).$$
Записываем $G^*_{t,k}$ и l^*.
3. В конце получилось ровно r траекторий, которые заканчиваются *в различных* состояниях при $t = T$. Выбираем $k^* = \arg\max_k G^*_{T,k}$.
4. Восстановление траектории из конца к началу. Полученная траектория называется *траекторией Витерби*.

Траектория Витерби

Время: $O(r^2 T)$

Для каждого $t \geqslant 2$ и для каждого $k = 1, ..., r$ перебираются все состояния предыдущего шага, при этом совершается $O(r^2)$ операций на каждом шаге.

Память: $O(rT)$

На каждом шаге для каждого состояния хранится оптимальное предыдущее состояния, т.е. $O(r)$ памяти на каждом шагу.

Задача

Пусть $Y = (Y_1, ..., Y_T)$ — наблюдаемая последовательность для некоторой скрытой марковской модели.

Для двух последовательностей скрытых состояний x и z (которые соответствуют некоторой траектории на графе развертки) определим величину $\rho(x, z) = \sum_{t=1}^{T} I\{x_t = z_t\}$ — количество общих состояний. Тогда $\mathrm{E}(\rho(X, z)|Y)$ имеет смысл среднего числа общих вершин у случайной траектории и заданной траектории z при условии, что наблюдается последовательность Y.

Доказать, что $\mathrm{E}(\rho(X, z)|Y)$ достигает максимума по z, если z — траектория forward-backward.